Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/28/eaba6883/DC1

Supplementary Materials for

Phylogenetic and physiological signals in metazoan fossil biomolecules

Jasmina Wiemann*, Jason M. Crawford, Derek E. G. Briggs

*Corresponding author. Email: jasmina.wiemann@yale.edu

Published 10 July 2020, *Sci. Adv.* **6**, eaba6883 (2020)
DOI: 10.1126/sciadv.aba6883

The PDF file includes:

Supplementary Text
Sections S1 to S8
Figs. S1 to S3

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/28/eaba6883/DC1)

Supplementary Data spreadsheet
Data files S1 to S3

Supplementary Text

A. Details on criteria for specimen selection

Specimens were selected for macroscopic evidence of 'carbonaceous' residues. Specimens with preserved organic matter were identified on the basis of criteria identified in Wiemann *et al.* (2018): we focused primarily on fossils that show a dark brownish or black discoloration which stands out against a light-colored sediment matrix. Light-colored sediments are usually low in sedimentary organic matter which could potentially 'contaminate' endogenous, fossil organic matter. Most fluvial, alluvial, aeolian, and shallow marine sediments fostered early diagenetic oxidative crosslinking as a prerequisite to biomolecule fossilization. We have complemented specimens known to preserve fossil organic matter based on analyses published in Wiemann et al. 2018a (*9*) (fossil vertebrate teeth, bones, eggshells) and Wiemann et al. 2018b (*13*) (fossil vertebrate eggshells) with additional specimens from suitable depositional environments. To further explore the limits of our methodological approach, pressure- and temperature-matured specimens from Paleozoic shales preserving macroscopic evidence for carbonaceous films were added to the data set. All specimen lithologies, ages and catalog numbers are recorded in the Supplementary Data spreadsheet under the tab 'Specimen data'. Catalog numbers can be used to access more detailed specimen data and photographs (as available) in the online catalog of the Yale Peabody Museum of Natural History Collections.

B. Method details

High resolution *in situ* Raman microspectroscopy setup and protocol

We used high-resolution *in situ* Raman microspectroscopy, a non-destructive method (*15, 16*), to analyze biomolecule fossilization products in 96 specimens of extant (n = 20) and representative fossil (n = 76) metazoans and their associated sediments (in 17 cases). From this dataset, we identified patterns in composition and biological significance by statistically analyzing different aspects of the spectrally encoded composition of biomolecules and their fossilization products. Fossil metazoan samples (n = 76) were selected for macroscopic evidence of dark brown- to black-colored carbonaceous films embedded in a lighter colored sediment matrix (available in n = 17 cases, Supplementary Materials, Section A). Preferred depositional environments included shallow marine, fluvial, alluvial, and aeolian oxidative settings that yield shallow marine limestones, light-colored sandstones, siltstones, and mudstones (following criteria identified in *9*). All samples and catalog numbers are listed in the Supplementary Data Spreadsheet under Specimen Data, and specimen photographs can be accessed via those numbers in the online catalog of the Yale Peabody Museum of Natural History. The analyzed fossil samples range in age from Cambrian to Tertiary, and provide taxonomic coverage of major animal groups with a notable fossil record, including sponges, corals, lophotrochozoans,

ecdysozoans, basal vertebrates, teleosts, chondrichthyans, amphibians, synapsids, and diapsids (marine reptiles, lizards, and archosaurs including non-avian and avian dinosaurs). All samples were individually surface-cleaned with 70% ethanol and immediately subjected to high resolution *in situ* Raman microspectroscopy.

Raman microspectroscopy was performed in the Department of Geology and Geophysics at Yale University using a Horiba LabRam HR800 with 532 nm excitation (20 mW at the sample surface). The spectra were obtained in LabSpec 5 software (spectra acquisition, standard spike removal). The scattered Raman light was detected by an electron multiplying charge-coupled device (EM-CCD) following dispersal with an 1800 grooves/mm grating and passed through a 200 μm slit (hole size 300 μm). The spectrometer was calibrated using the first order Si band at 520.7 cm^{-1}. Ten spectra were accumulated in the 500-2000 cm^{-1} region for 10s exposure each, at 32x magnification. All spectra were analyzed in SpectraGryph 1.2 spectroscopic software (adaptive baseline, 30%, no offset, minimally smoothed through rectangular averaging over an interval of 4 points). Spectra were plotted, and individual Raman bands in the spectra were identified, through an automated peak search function (see Supplementary Data Spreadsheet, Band Sets 1-4) in SpectraGryph 1.2. Spectral averaging, as performed for analyses shown in Figs. 1a and 2b (bar charts), was also performed in SpectraGryph 1.2 (spectral averaging function, mean values).

3-D ChemoSpace Plot and Discriminant Analysis: Organic matter through time

We selected 35 Raman bands from the organic fingerprint region (500-1800 cm^{-1} Raman Shift) to characterize the change in total organic composition (total modern and fossil samples, n = 93) evident in averaged spectra for eggshells (modern, n = 4; fossil, n = 14), bones (modern, n = 4; fossil, n = 19), teeth/enamel scales (modern, n = 3; fossil n = 14), vertebrate soft tissues (modern, n = 3; fossil, n = 9), biomineralized invertebrate tissues (modern, n = 3; fossil, n = 9), and invertebrate soft tissues (modern, n = 3; fossil, n = 11) during fossilization (Supplementary Materials, Section B). Analyzing selected band intensities over whole-spectral data avoids potential curvilinear distortion of signals in vibrational data analyzed through ordination methods, such as a Principal Components Analysis (ChemoSpace) (*17*). The selected bands characterize organic functional groups present in extant and fossil sample spectra and are assigned in the Supplementary Data Spreadsheet, Band Set 1: 540, 630, 660, 680, 723, 750, 776, 830, 840, 880, 923, 956, 978, 1003, 1060, 1075, 1107, 1171, 1189, 1205, 1242, 1281, 1294, 1349, 1412, 1431, 1464, 1483, 1528, 1582, 1643, 1663, 1695, 1752, 1775 $\pm$ 2 cm^{-1} (systematic uncertainty posed by the instrument setup) Raman Shift. Relative spectral intensities were averaged for taxonomic categories in SpectraGryph 1.2 spectroscopic software, transformed into a variance-covariance matrix, and subjected to a ChemoSpace Principal Component Analysis in PAST 3 software. The data for each averaged taxonomic category in the resulting 2-D ChemoSpace were complemented, through an automated peak search function in SpectraGryph 1.2, with the total number of different, significant peaks and bands representing distinct organic functional groups. We exported {x, y}-coordinates for the position of each taxonomic category

in the ChemoSpace, and treated the corresponding number of functional groups as {z}-coordinates (all data are provided in the Supplementary Data Spreadsheet, 3-D ChemoSpace). Based on these Cartesian coordinates a 3-D plot (Fig. 1a) illustrating the diversity in organic functions versus the distinctness of the composition of each sample (fossil or modern) was generated; modern sample data were colored in blue and fossil data in orange. Modern, unaltered biomolecules form an almost {x, y}-parallel plateau of high functional diversity and cover a different {x, y}-area than fossil samples in this 3-D ChemoSpace. During fossilization the total organic tissue composition converges on a functionally simpler composition (shifts from its modern composition along the red eigenvector representing an increase in glycoxidation and lipoxidation markers), as the diversity of distinct organic groups falls to a third in fossil tissues (see Fig. 1a, compare Fig. S1). Modern, remodeled and vascularized tissues, such as teeth (dentine) and bones, are known to record crosslinks formed *in vivo* resulting from metabolic stress (*18-21*), and are therefore shifted in the ChemoSpace towards increased amounts of AGEs and ALEs.

Using the same tissue category-averaged data set as that in the 2-D ChemoSpace, with the addition of a single parameter assigning a modern or fossil identity to each tissue category, this data matrix was subjected to a Discriminant Analysis in PAST 3 software. Fossil tissues are represented by orange dots, while modern tissues are represented by blue dots. Key discriminant features between the two sample sets are listed in the Discriminant Analysis.

A novel proxy to quantify diagenetic alteration: the *trans*-/*cis*-amide ratio

Diagenetic alteration was assessed through the degree of protein transformation in fossil metazoans. Proteins crosslink with lipid- or sugar-derived RCS contributing to the polymerization process involved in the fossilization of these structural biomolecules. Whereas initial crosslinking reactions affect only suitable amino acid residues (Fig. 2c), advanced crosslinking involves the peptide skeleton which is characterized by its dominant peptide bonds (*trans*-amides (A_p), Fig. 4e) (reviewed in *12*). Advanced crosslinking tends to produce heterocycles through condensation reactions which yield characteristic *cis*-amide (A_d) signals in the Raman spectra of fossil animal samples. Potential *cis*-amides in fresh proteins, lipids, and sugars yield a very weak Raman signal, so the strong signal intensity in fossil organic matter indicates clearly that detected *cis*-amides are products of diagenetic transformation. The ratio of intact peptide bonds (*trans*-amides) to diagenetic reaction products of peptides (*cis*-amides) provides an assessment of the degree of diagenetic alteration of structural biomolecules in fossil animal samples (Fig. 2b). Relative signal intensities for the bands at 1685 cm^{-1}, representing *trans*-amides, and 1702 cm^{-1}, representing *cis*-amides, were extracted from averaged spectra for sample categories: eggshells (n = 14), teeth (n = 14), bones (n = 19), vertebrate soft tissues (n = 9), biomineralized (n = 9) and non-biomineralized invertebrate tissues (n = 11) of fossil metazoan sample spectra (n = 76), and the ratio of *trans*-/*cis*-amides was calculated and plotted (Fig. 2a, bar charts). Values of (A_p/A_d) >1 indicate that more peptide bonds are intact than are

diagenetically altered, allowing the preservation of the organic phase in different samples to be compared.

Discriminant analysis: Fossil versus sediment organic matter

The possibility of contamination with exogenous organic matter is a common concern where signals indicate intact peptide bonds (Fig. 2a, b). We test this by analyzing organic matter in sediment associated with the fossil samples. We determined signal intensities at 24 band positions characterizing the molecular composition of fossil organic matter (compare to the spectra in Fig. 2a) in the Raman spectra of all metazoan fossils (n = 76) and associated sediment samples (n = 17). The selected bands correspond to a range of different organic functional groups which are assigned in the Supplementary Data Spreadsheet, Band Set 2: 510, 574, 630, 660, 776, 843, 879, 910, 1009, 1032, 1115, 1187, 1231, 1278, 1345, 1408, 1450, 1489, 1553, 1572, 1594, 1702, 1774 $\pm$ 2 cm^{-1} (systematic uncertainty posed by the instrument setup) Raman Shift. Relative intensities in SpectraGryph 1.2 spectroscopic software were transformed into a variance-covariance matrix and labeled as fossil or sediment. These data were subjected to a Discriminant Analysis in PAST 3 software, and the resulting biplot was exported. Discriminant factors between fossil and sediment were identified based on their Raman band assignments. *Trans*-amides (= peptide bonds), thioethers and S-heterocycles are characteristic of fossil animal organic matter, while peroxidized aromatics and aliphatics are characteristic of sediment organic matter. The two sample groups have distinct compositions, and the peptide bonds are not a product of sediment or sample contamination.

Organo-sulfur species abundance in fresh and fossil metazoan tissues

Reductive sulfurization can yield thioethers, and may represent an alternative process to oxidative crosslinking. Sulfurization acts under highly reducing conditions, in contrast to oxidative crosslinking. We conducted a series of analyses (Figs. S2, S3) to test for the incorporation of environmental sulfur into geopolymers, and to characterize the potential effect of sulfurization on our interpretation of S-heterocycles in metazoan fossil organic matter from primarily oxidative settings.

Normalized and averaged Raman spectra of all investigated fossil and extant tissue types (Fig. 2a, compare to Fig. S1) were plotted in the range of 500-700 cm^{-1} Raman shift, and were paired based on the abundance and spectral range signatures of organo-sulfur species. Fresh and fossil tissues were analysed separately. Diagnostic organo-sulfur signatures can be found in the range between 510 and 690 cm^{-1} Raman shift. Organo-sulfur species show a comparable composition and relative abundance in bone and dental tissues, in vertebrate and invertebrate soft tissues, and in vertebrate eggshells and biomineralized invertebrates (Fig. S2). Ranges in the relative abundance of certain organo-sulfur species (= functional groups corresponding to the specific Raman shift) were colored in orange for bone and dental tissues, in pink for vertebrate and invertebrate soft tissues, and in blue for vertebrate eggshells and biomineralized

invertebrates. Correspondence in the abundance of organo-S correspond between fossil and extant metazoan tissues argues against the incorporation of environmental sulfur through reductive sulfurization.

Distribution of organo-sulfur species in fossil and sedimentary organic matter

To test for any significant contribution of sulfurization processes to the detected S-heterocycles, we further analyzed the similarity in the organo-sulfur species composition of fossil and sedimentary organic matter. Relative intensities were extracted for all spectra of fossil and sediment samples (n = 93) at 13 band positions characterizing the detailed composition of organo-sulfur species (in SpectraGryph 1.2 spectroscopic software). Organo-sulfur species yield diagnostic signatures between 500 and 700 cm^{-1} Raman shift. The selected band positions are assigned in the Supplementary Data Spreadsheet, Band Set 3 and include: 510, 520, 530, 560, 570, 580, 615, 630, 645, 655, 665, 675, 695 $\pm$ 2 cm^{-1} (systematic uncertainty posed by the instrument setup) Raman Shift. Relative intensities in SpectraGryph 1.2 spectroscopic software were transformed into a variance-covariance matrix. This matrix contained a total of n = 93 fossil metazoans and sediment samples, and was subjected to a ChemoSpace Principal Component Analysis and a separate Discriminant Analysis in PAST 3 software.

For the PCA ChemoSpace (Fig. S3a), fossil metazoan data points were colored in blue, while sediments were colored in grey. Correspondingly colored convex hulls show that even a substantially reduced data set using only organo-sulfur signatures is almost sufficient to separate the clusters of fossil and sediment organic matter. The full band set analyzed (Fig. 2d) has an increased discriminatory power, however, the ChemoSpace (Fig. S3a) shows that there are significant differences in the organo-sulfur composition in fossil and sediment organic matter (see separation across PC 1, 79%). Eigenvectors were plotted, and sediment organic matter appears to differ from fossil metazoan organic matter primarily in more abundant sulfides, disulfides, and thiols (colored in orange), while fossil organic matter contains more abundant thioethers and S-heterocycles (colored in blue).

The Discriminant analysis (Fig. S3b) analyzes the same taxon-character matrix used in the PCA ChemoSpace (Fig. S3a), with the addition of a sample identifier (fossil metazoan *versus* sedimentary organic matter). Fossil (blue) and sediment (grey) organic matter data points separate along the discriminant vectors (red arrows). Discriminant vectors correspond to the eigenvectors of the PCA ChemoSpace. Organo-sulfur species are distinct in fossil and sedimentary organic matter, arguing against the incorporation of environmental sulfur via reductive sulfurization processes.

2-D ChemoSpace PCA and Discriminant Analysis: Tissue type clustering

A discriminant analysis of fossil samples was carried out to explore the possibility and nature of a tissue-specific signature (Figs. 1a, 2a). The taxa (n = 76) sampled in the variance-covariance matrix used in the Discriminant Analysis of fossils and sediments (Fig. 2d) were

grouped into biomineralized and non-biomineralized categories. Including a biomineralization identifier, the resulting data matrix was subjected to a Discriminant Analysis in PAST 3 software incorporating this additional parameter, and the resulting biplot was exported (Fig. 3a). Features characteristic of biomineralized animal samples were identified as abundant thiols, S-heterocycles, peptide bonds, N-, and O-heterocycles. These functional groups can be categorized as peptide bonds and chelating ligands (= functional groups prone to engaging in intermolecular interactions with a mineral phase). Features characteristic of non-biomineralized animal samples, in contrast, are abundant diagenetic amides (=*cis*-amides), aliphatics, lactones (O-heterocycles common in peroxidized organic matter) and esters. These functional groups indicate that fossil organic matter has been altered by degrading peroxidation.

The 24 Raman band intensities selected for fossil metazoans (n = 76) were subjected to a ChemoSpace Principal Component Analysis in PAST 3 software without the parameter discriminating biomineralized and non-biomineralized samples. PC 1 (56%) and PC 2 (22%) separate samples characterized by features identified by the corresponding Discriminant Analysis (Fig. 3a).

A second analysis selected vertebrate hard tissues (n = 41) from the matrix of 24 Raman band intensities for fossil metazoan spectra (n = 76) (Figs. 2c, 3a). Three categories were identified: eggshells (n = 14), bones (n = 19), and teeth excluding conodonts (n = 8) (Fig. 2a). The data matrix incorporating these additional parameters was subjected to a Discriminant Analysis in PAST 3 software, and the resulting plot was exported. Features characteristic of fossil eggshells were identified as abundant aliphatics, lactones, esters, peptide bonds, thioethers, and S-heterocycles whereas features characteristic of bone are abundant aliphatics, N-heterocycles, ketones, and lactones. Teeth contain more abundant aromatics, diagenetic amides (= *cis*-amides), and thiols.

The 24 Raman band intensities selected for fossil metazoans (n = 76) were subjected to a ChemoSpace Principal Component Analysis in PAST 3 software without distinguishing between eggshell, bone and teeth. PC 1 (43%) and PC 2 (31%) separate samples characterized by features identified by the corresponding Discriminant Analysis (Fig. 3b). Convex hulls show no overlap in the molecular composition of calcite- and apatite-biomineralized fossil samples (Fig. 3b).

Spectral Cluster Analysis: Preserved phylogenetic signals in different tissue types

The composition of protein fossilization products (PFPs) varies not only with original differences but also with diagenesis, which may reduce a potential phylogenetic signal. All samples were screened for PFPs: relative spectral intensities at 550 cm^{-1} (S-heterocycles) Raman shift and 1580 cm^{-1} (N-heterocycles) Raman shifts were determined for samples of every taxon in the data matrix generated for the ChemoSpace analysis (Fig. 3a). S- and N-heterocycles are transformation products of protein crosslinking (Figs. 2b, 4e, S1). The ratio of S- and N-heterocycles reflects the original amino acid composition and retains phylogenetic information (Fig. S1). Outlying values (determined through a quartile-based outlier analysis individually applied to each tissue category) of the S-heterocycle/N-heterocycle ratio ([-C-S-]/[-C-N-]) from

each sample category were omitted from the data set. Within each tissue category, samples of identical or directly comparable lithologies were selected to avoid signal distortion through minute differences in the mode of organic matter preservation (lithologies for all samples are listed in the Supplementary Data Spreadsheet, Specimen Data). The remaining taphonomically comparable fossil samples (highlighted in the Supplementary Data Spreadsheet, Specimen Data) – eggshells (n = 8), teeth (n = 6), bones (n = 10), biomineralized invertebrate samples (n = 7), and non-biomineralized invertebrate samples (n = 6) – were used to determine signal intensities at 20 band positions characterizing the specific composition of amino acid-derived N- and S-heterocycles (Figs. 2c, S1) in their preserved geopolymers. Proteins are the only crosslinking type of compounds that contain phylogenetic information (contrary to lipid- and sugar-derived RCS). To extract this signal, a refined band set was created based on the concepts presented in Figs. 2c, 4f, and S1. The selected bands correspond to a range of functional groups present in different amino acid-derived crosslinks which are assigned in the Supplementary Data Spreadsheet, Band Set 4: 513, 525, 540, 560, 576, 584, 605, 628, 662, 670, 691, 1398, 1432, 1455, 1476, 1516, 1577, 1589, 1600, 1650 $\pm$ 2 cm^{-1} (systematic uncertainty posed by the instrument setup) Raman shift. Relative intensities in SpectraGryph 1.2 spectroscopic software were transformed into a variance-covariance matrix.

Using this optimized taxon-character matrix, we ran two different types of cluster analyses: a cross-tissue analysis that finds a monophyletic clade of vertebrates nested within invertebrates, and individual cluster analyses for each tissue type. For the cross-tissue cluster analysis, all pre-screened samples suitable for phylogenetic analyses were used, except for vertebrate eggshells, the siliceous sponge, and *Lithostrotion*. These samples were excluded due to the (potential) lack of 'homologous' proteinaceous matter in their tissues. Eggshell, sponge and *Lithostrotion* represent tissues which contain primarily proteins acting as biomineral chelates and templates, but lack any conserved proteinaceous structures. The resulting data set of n = 27 fossil metazoan tissues was subjected to a hierarchical cluster analysis in PAST 3 software, and a rooted topology (*Rho*, one-way, unconstrained) was generated (Fig. 4a).

For the tissue type phylogenetic clustering, tissue-specific taxon-character matrices were treated as separate categories, and subjected to a hierarchical cluster analysis in PAST 3 software. A rooted topology was generated for each category, and one-way *Rho* was selected as the mode of clustering; except for the tree rooting, the analysis was not constrained. The resulting topologies were compared to published consensus trees, and a measure of correspondence was calculated by assessing each node for correct resolution. The phylogeny used for comparison with the topology generated from fossil organic matter in dental tissues is a consensus tree based on Simoes et al. 2018 (*22*, Diapsida) and Meyer and Zardoya 2003 (*23*, Vertebrata). The tree compared to the topology generated from fossil organic matter in eggshells is based on Brusatte et al. 2014 (*24*); the tree compared to the topology generated from biomineralized and non-biomineralized invertebrates is based on Glenner et al. 2004 (*25*). The tree used to assess the topology generated from fossil organic matter in bone is a consensus phylogeny based on Gauthier et al. 1986 (*26*, Dinosauria), Simoes et al. 2018 (*22*, Diapsida), and

Meyer and Zardoya 2003 (*23*, Vertebrata). Topologies were exported and redrawn, and nodes were colored according to their degree of correspondence with published topologies (Fig. 4a, c, d, e): correctly resolved nodes are colored in yellow, incorrectly resolved nodes in red, and nodes that are currently unresolved (polytomies) in published consensus trees are colored in grey.

Using the taxon-character matrix generated from Band set 4 (Supplementary Materials) for fossil bone tissues, a tissue-specific ChemoSpace analysis (Fig. 4b) was run in PAST 3 (Principal Component Analysis, variance-covariance). Most of the variation in the data set of N- and S-crosslinks in bone can be explained based on metabolic sorting: taxa with few *in vivo* formed metabolic crosslinks plot in the y-positive quadrants of the ChemoSpace, while taxa with high metabolic rates plot in the y-negative quadrants. Based on the literature on fossil metabolic rates, data points were colored in orange for taxa with a metabolic rate > 1 ml O_2 h^{-1} $g^{-0.67}$, or blue for taxa with a metabolic rate < 1 ml O_2 h^{-1} $g^{-0.67}$. Bone and dental soft tissues are known to record metabolic stress in the form of protein modifications which seem to overwrite some of the phylogenetic information (*12, 18, 19, 20, 21*).

The extracted S-/N-heterocycle ratios were averaged for all categories to investigate why the accuracy of the phylogenetic signal differed between them. While the [-C-S-]/[-C-N-] ratios for fossil tissues with a strong phylogenetic signal range from 0.11-0.52, ratios for bones and teeth (0.98-1.40) show increased values of S-heterocycles. Bones and teeth are the only vascular tissues in this data set. Since vascular tissues are remodeled, and thereby directly affected by the vertebrate metabolism, phylogenetic signals in their structural proteins may be overprinted *in vivo* by a metabolic signal (*11, 20, 21*).

C. Rationale behind Raman band selections

High-resolution *in situ* Raman microspectroscopy is a highly efficient and reliable non-destructive method to characterize trends in the molecular composition of fossils (*9, 13*). Raman spectral data can be highly informative when analyzed by means of multivariate statistics (*9, 13*), but most commonly require band-selective conversion into a taxon-character matrix. When analyzed as whole spectral data, broad bands – which are especially common in fossil materials containing numerous degradation products of similar composition – may induce curvilinear distortion in the ChemoSpace (*17*). When assessing a ChemoSpace for biologically informative sample clustering, such distortion would result in a reduced signal separation (*17*). To avoid such effects and achieve the best possible representation of potential biological signals in a fossil ChemoSpace, informative Raman band intensities should be selected (*17*). Since Raman spectroscopy records inorganic mineral as well as organic signals in fossils, analyses of signals in the organic phase should omit mineral peaks (*9, 13*).

In this study we used four different band sets to address two different questions. Band Set 1 was used to highlight how the composition of fresh and fossil animal tissue differs, and therefore covers Raman bands and peaks characterizing unaltered proteins, lipids, and sugars, as well as prominent bands characteristic of fossil organic matter (compare *9*). The band set was generated by plotting Raman spectra for unaltered biomolecules as well as fossil metazoan organic matter

(see Fig. 2a for the variation in the latter). Note that the labels shown in Fig. 2a characterize the band shifts of all superimposed fossil metazoan spectra which together yield the averaged spectra shown (solid black line = averaged fossil metazoan, dotted red and blue lines = averaged fossil invertebrate and vertebrate). All selected Raman bands in Band Set 1 and their detailed assignments, as well as the Raman shift uncertainty posed by the instrument, are recorded in the Supplementary Data Spreadsheet, Band Set 1.

Band Set 2 was used to tease out tissue-specific compositional heterogeneities in fossil organic matter – thus only Raman bands characterizing fossil organic matter are included. Bands related to fresh, unaltered biomolecules are omitted here, as our analyses (Figs. 1 and 2) demonstrate that they are transformed chemically during fossilization. Band Set 2 is based on Raman bands present in fossil vertebrates and invertebrates, and band positions have been selected based on the band set in Wiemann et al. 2018a (*9*). Band Set 2 was used to distinguish fossil animals and sediment organic matter (Fig. 2d), and was subsequently analyzed for retained tissue type signals (Figs. 3, 4a-d). The complete Band Set 2, detailed band assignments, and the systematic instrument uncertainty are recorded in the Supplementary Data Spreadsheet, Band Set 2.

Band Set 3 was used to test for the distribution of organo-sulfur species in fossil metazoan and sedimentary organic matter. To assess whether sulfurization processes contribute to the detected thioethers and S-heterocycles, a band set characterizing details of the organo-S composition was generated based on the presence of peaks in the spectral range between 500 and 700 cm^{-1} Raman shift. Normalized intensities were selected for a set of n=13 bands present in fossil metazoan and sediment organic matter. The complete Band Set 3, detailed band assignments, and the systematic instrument uncertainty are documented in the Supplementary Data Spreadsheet, Band Set 3.

Band Set 4 focuses on the specific composition of amino acid-derived crosslinks to characterize the quality of a preserved phylogenetic signal. The bands are selected based on present peaks in the pre-selected fossil vertebrates and invertebrates (see Material and Methods for outlier analysis and lithology). A set of n=20 bands characterizing amino acid-derived crosslinks (based on *9*) was selected from the spectral range of 500-1800 cm^{-1} Raman shift. The complete band set 4, detailed band assignments, and the systematic instrument uncertainty are documented in the Supplementary Data Spreadsheet, Band Set 4.

D. Additional details on diagenetic reaction schemes

Analyses of spectral data on the composition of fresh and fossil animal tissues indicate that a suite of chemical transformations occurred during the fossilization process. The ChemoSpace PCA (Fig. 1a) illustrates how the shift in molecular composition from fresh to fossil tissues follows a trajectory of glycoxidation and lipoxidation processes, while the Discriminant analysis associated with Fig. 1a adds more details on specific transformations, such as the conversion of amino acid thiols into thioethers and S-heterocycles, and the conversion of amines in amino acids into N-heterocycles. Fig. S1 shows a spectral net enrichment plot for average extant versus

average fossil animal composition. All spectra obtained for extant animal tissues (data set for Fig. 1a) were averaged, as were all spectra obtained for fossils (used in Fig. 1a, plotted in Fig. 2a) were averaged. The averaged extant animal spectrum was subtracted from the averaged fossil animal spectrum, and plotted. Excursions into negative values on the y-axis represent a net depletion in these molecular functions during fossilization, while peaks in positive values on the y-axis represent a net enrichment during fossilization. Peaks, both negative and positive, can be interpreted just as a regular Raman spectrum, and detailed band assignments can be found in the Supplementary Data Spreadsheet, Band Set 2.

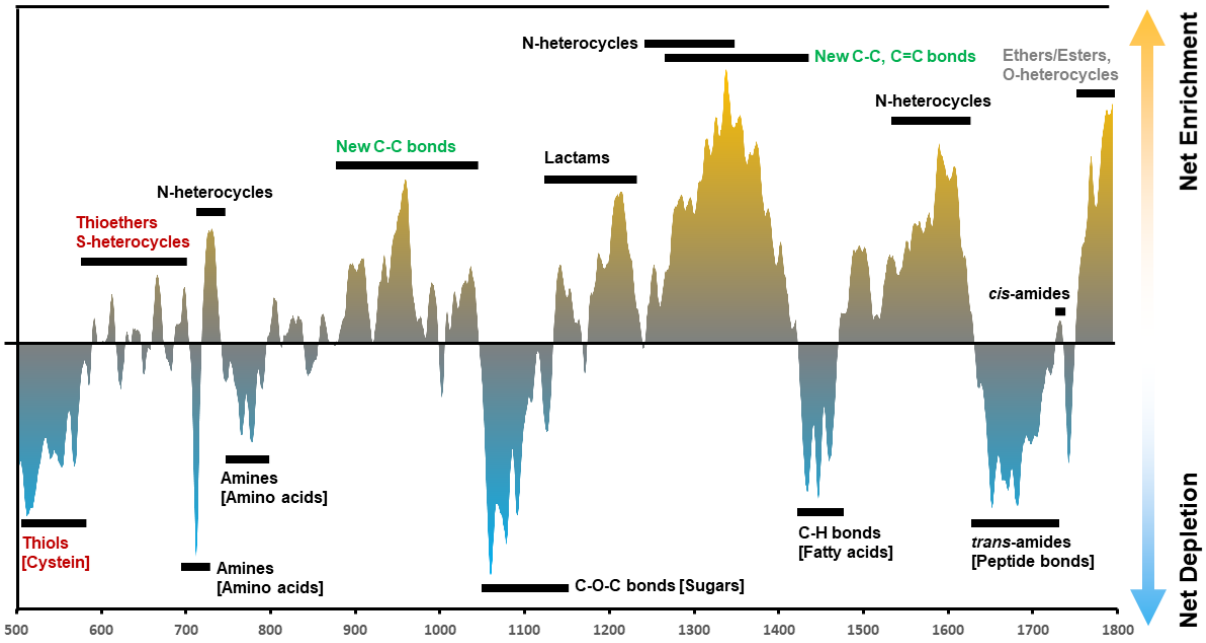

Fig. S1: Net enrichment plot of averaged fresh and fossil animal tissues. Peaks on the enrichment side indicate structures newly formed during fossilization (= reaction products), while peaks on the depletion side identify structures lost during fossilization (= reaction educts). Labels on the positive and negative peaks offer general band affinities. The trends shown are consistent with glycoxidation and lipoxidation as key processes during the fossilization of animal biomolecules. Thiol- and amine-bearing amino acids crosslink with lipid- (fatty acid) or sugar-derived Reactive Carbonyl Species with the formation of new C-C bonds (electrophilic additions) to yield N-, O-, S-heterocyclic polymers.

The net enrichment plot shown in Fig. S1 shows that chemical transformations during the fossilization of biomolecules consume thiols, amines, saccharides, fatty acids, and peptide bonds. Fossilization reactions yield newly formed C-C/C=C bonds, thioethers, ethers, esters, and *cis*-amides associated with N-, O-, S-heterocycles. These reaction educts and fossilization products are consistent with glycoxidation and lipoxidation reactions (reviewed in *12*).

E. Evidence against sulfurization as crosslinking mechanism

Reductive sulfurization is a well-studied early diagenetic process yielding thioethers and other sulfur-crosslinks primarily in degradation products of lipids. Thus sulfurization may

represent an alternative process to oxidative crosslinking that occurs under highly reducing conditions (*10*). Since our study aims to evaluate preserved biological signals in fossil metazoan organic matter, endogeneity of the interpreted N-, O-, S-heterocyclic polymers is a fundamental requirement. We conducted a series of analyses (Figs. S2, S3) to test for the incorporation of environmental sulfur into geopolymers, and to characterize the potential effect of sulfurization on our interpretation of S-heterocycles in metazoan fossil organic matter from primarily oxidative settings.

Reductive sulfurization incorporates environmental organo-sulfur into early degradation products of lipids cleaved via hydrolysis, which occurs exclusively in highly reducing settings (*10*). All samples used in our study, except for the four Burgess shale specimens, come from oxidative depositional settings, as evidenced by the host rock sedimentology (see Supplementary Data Spreadsheet, Specimen Data). Reductive sulfurization does not occur under oxidative conditions. Burgess Shale specimens are preserved as carbonaceous films with associated clay minerals on the surface. Clay minerals form abundant Reactive Oxygen Species along the catalytic surfaces of their different layers, generating locally oxidative conditions (*28*), which would not be conducive to reductive sulfurization.

We analysed the relative abundance of organo-sulfur species in fresh and fossil metazoan organic matter. Normalized and averaged Raman spectra of all investigated fossil and extant tissue types (Fig. 2a, compare to Fig. S1) were plotted in the range of 500-700 cm^{-1} Raman shift, and were paired based on the abundance and spectral range signatures of organo-sulfur species. Fresh and fossil tissues were analysed separately. Diagnostic organo-sulfur signatures occur in the range between 510 and 690 cm^{-1}. Organo-sulfur species show a comparable composition and relative abundance in bone and dental tissues, in vertebrate and invertebrate soft tissues, and in vertebrate eggshells and biomineralized invertebrates (Fig. S2).

Samples were grouped by tissue types to allow us to explore tissue-specific differences. If sulfurization contributed to the formation of the thioethers and S-heterocycles detected, fossil organic matter would likely contain more organo-sulfur than its fresh tissue analogue (inferred from *10*). In addition the distribution of organo-sulfur abundances in fossils would not correspond to that in fresh tissues (inferred from *10*). However, Fig. S2 shows that all analyzed fossils contain less organic sulfur than their extant tissue analogues, and that the abundance of detected organo-sulfur species in fossil organic matter scales with that in analogous fresh tissues. Such a correspondence between fresh and fossil organo-sulfur abundance does not support the diagenetic incorporation of environmental sulfur.

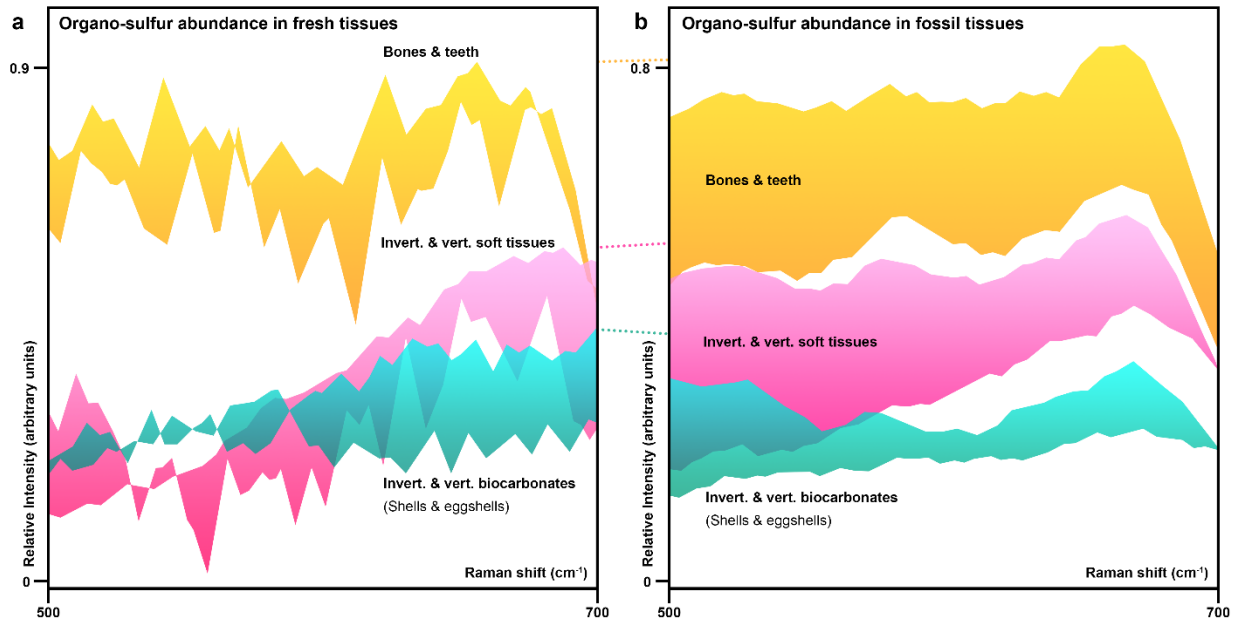

Fig. S2: Relative abundance (represented by normalized intensity) of different organo-sulfur species in fresh and fossil tissues over a Raman range from 500 – 700 cm^{-1}. Tissue types are grouped by relative amounts of organo-sulfur: blue band = biomineralized invertebrates and vertebrate eggshells, pink band = invertebrate and vertebrate soft tissues, and orange band = vertebrate bones and teeth. Note that the scale on the y-axis differs between the two plots, showing generally reduced amounts of organo-sulfur in fossil organic matter. The dotted lines between the two plots correspond in color to the tissue type groupings, and show the difference in maximum intensity of organo-sulfur in fresh and fossil tissues.

Reductive sulfurization affects both animals and sedimentary organic matter (inferred from *10*). Lipid degradation products are the primary target (*10*) and they are present in all life forms (i.e., in cell membranes), including the likely sources of organic compounds in sediment, such as plants and microbes, as well as metazoans. Therefore, sulfurization should leave similar organo-sulfur species in fossil and sediment organic matter (inferred from *10*). To test for any significant contribution of sulfurization processes to the detected S-heterocycles, we analyzed the similarity in the organo-sulfur species composition of fossil and sedimentary organic matter. Relative intensities were extracted for all spectra of fossil and sediment samples (n = 93) at 13 band positions characterizing the detailed composition of organo-sulfur species (in SpectraGryph 1.2). Organo-sulfur species yield diagnostic signatures between 500 and 700 cm^{-1} Raman shift. The selected band positions are assigned in the Supplementary Data Band Set 3. Relative intensities in SpectraGryph 1.2 were transformed into a variance-covariance matrix of n = 93 fossil metazoans and sediment samples, and was subjected to a ChemoSpace Principal Component Analysis and a separate Discriminant Analysis in PAST 3 (Fig. S3).

The full band set analyzed (Fig. 2d) has greater discriminatory power, but the ChemoSpace (Fig. S3a) shows that even the reduced data set of organo-sulfur signatures reveals significant differences in the organo-sulfur composition in fossil and sediment organic matter (see

separation across PC 1, 79%). Eigenvectors indicate that sediment organic matter differs from fossil metazoan organic matter primarily in more abundant sulfides, disulfides, and thiols, whereas fossil organic matter contains more abundant thioethers and S-heterocycles.

The Discriminant analysis (Fig. S3b) analyzes the same taxon-character matrix, with the addition of a sample identifier (fossil metazoan *versus* sedimentary organic matter). Organo-sulfur species are distinct in fossil and sedimentary organic matter, arguing against the incorporation of environmental sulfur via reductive sulfurization processes. Such a substantial difference in organo-S composition would not be expected if it were the result of sulfurization, which would affect both sediment and fossil organic matter. We conclude that sulfurization was not a major contributor to the formation of the N-, O-, S-heterocyclic polymers detected.

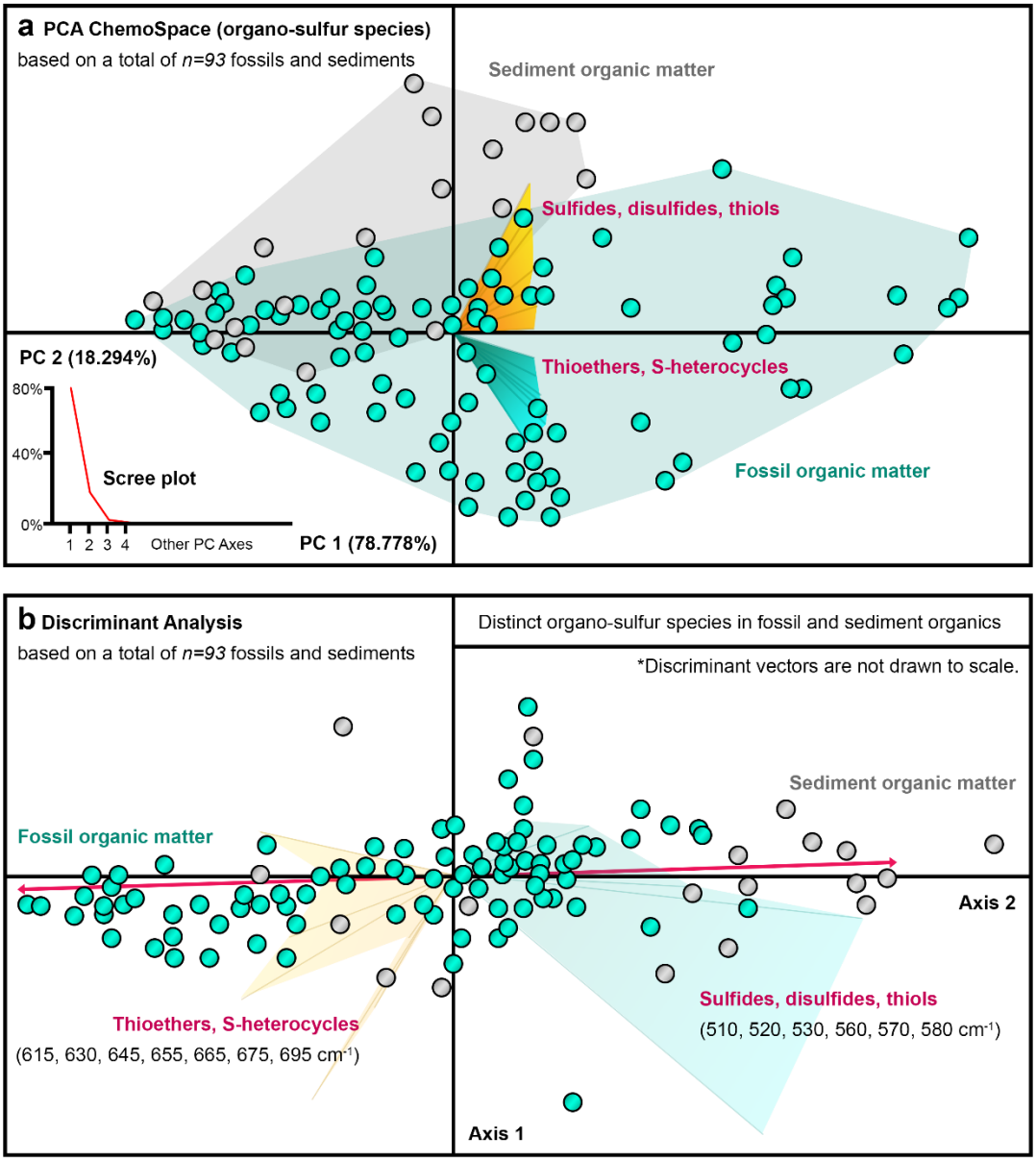

Fig. S3: PCA ChemoSpace and Discriminant Analysis of organo-sulfur species in fossil and sedimentary organic matter. Raman spectral bands representative of S-crosslinks (n=13) discriminate between fossil and sediment organic matter. Fossil organic matter is represented by turquoise dots, whereas sediment organic matter is represented by grey dots. **a** PCA ChemoSpace of n = 93 fossils and sediments. Eigenvector trajectories for thioethers and S-heterocycles (blue vector area) place fossil samples mainly among negative values of PC1, whereas sulfides, disulfides, and thiols (orange) place sediment samples mainly among positive values of PC1. The scree plot shows that most of the variability in the organo-sulfur species distribution reflects the source of the sample (either fossil or sediment organic matter) (PC1 79%). **b** Discriminant analysis of n = 93 fossils and sediments. Discriminant vector areas are not drawn to scale. The clusters of fossil and sediment samples are separate and different types organo-S species discriminate between fossil and sedimentary organic matter. If sulfurization impacted these fossils, both fossil and sedimentary organic matter would be affected equally and no distinct organo-sulfur signatures would be expected.

A plot of net enrichment during fossilization (Fig. S1) reveals reaction educts and products: amino acid residues (specifically those containing amines and thiols), reducing sugars, and lipids are consumed, and crosslinks form. These crosslinks include thioethers, S-heterocycles, newly formed C-C bonds, N-heterocycles, carbonyls, carboxyls, and O-heterocycles. Oxidative crosslinking (*sensu* Advanced Glycoxidation and Lipoxidation) corresponds to this reaction scheme (*12*). Sulfurization can generate thioethers under suitable conditions, but fails to generate N-, O-, and S-heterocycles, as well as new C-C bonds, carbonyls and carboxyls (*10*).

F. Biological signals in the ChemoSpace

When analyzing spectral data obtained from fossils, a suite of different signals can be extracted by assessing the ChemoSpace PCA eigenvectors or PC loadings. Cross-comparisons of mineral-inclusive spectral band sets for fossils from depositional environments, and therefore different chemomilieus/redox milieus, tend to cluster taxa on PC1 based on their sedimentary environments. In order to focus on biological signals, samples from comparable depositional environments are selected, and only organic bands are analysed. Band Sets 2 and 4 (Supplementary Data Spreadsheet) allowed us to test for the presence of paleobiologically informative signals in fossils. Such biological signals encompass a biomineralization signal, a tissue type signal, and a phylogenetic signal. Assessment of the fidelity of the phylogenetic signal revealed that *in vivo* remodeled tissues may retain additional information reflecting a metabolic signal (see Fig. 4b). These different biological signals can be characterized by targeting different categories and features of fossil organic matter. When comparing biomineralized and non-biomineralized fossil animal tissues, PC1 is shown to sort samples based on the abundance of chelating and coordinating ligands in the fossil organic matter (Fig. 3a). When comparing biomineralizers and non-biomineralizers, the degradation products of *in vivo* chitinous tissues are separated from those of lipid-rich tissues (see sample sorting within each cluster: Fig. 3a) based on the relative abundance of AGEs (Advanced glycoxidation end products) and ALEs (Advanced lipoxidation end products). This separation distinguishes between chitinous invertebrate and non-chitinous vertebrate structural tissues (*34*). When comparing biomineralized vertebrate samples from comparable depositional environments, PC1 sorts samples based on the biomineral-specific composition of the organic phase (applied in *35*), while PC2 reflects some phylogenetic input. In this case phylogenetic signals are the least prominent, as they rely on minute differences in the amino acid-specific reaction products of crosslinking. If all the samples were included in the ChemoSpace the result would be a separation where PC 1 represents a taphonomic signal, PC 2 a biomineralization and tissue type signal, and PC 3 a phylogenetic signal (suggested by the PC Loadings).

The PC loadings for the analyses shown in the Figs. 3a and b can be accessed in the Source Data PAST-files for those figures. Loadings/eigenvector interpretations are included in Figs. 3a, b.

G. Subsampling and outlier analysis for the phylogenetic clustering

The phylogenetic signal in our spectral data for fossil animals relies largely on differences in the composition of thioethers, S- and N-heterocycles (see main text). To avoid any sample clustering based on differences in taphonomy, biomineralization or other tissue type signals, samples from identical or at least very similar lithologies were analyzed individually within tissue categories. In addition to this subsampling of the original fossil animal data set (n=76), an outlier analysis was performed to identify any significant alterations of the fossil organic phase. Specimen lithologies are listed in the Supplementary Data Spreadsheet, Specimen data. The outlier analysis focuses on the phylogenetically informative S- and N-heterocycle signatures (Fig. 4). S- and N-heterocycle signal intensities were selected at 530 cm^{-1} (-C-S-) and 1580 cm^{-1} Raman shift (-C-N-) respectively, and a ratio was calculated for each sample. Within each tissue type, these ratios were subjected to an outlier analysis in Microsoft Excel. The first and third quartiles were calculated for ratios in each tissue category, and subtraction of these values yielded the interquartile range. Lower and upper limits were calculated using the following equations, and any individual [-C-S-]/[-C-N-] ratio within a tissue category was assessed to determine whether its value falls within the range or represents an outlier.

Lower limit equation

$$Lower\ limit = [1st\ Quartile] - 1.5 \times [Interquartile\ range]$$

Upper limit equation

$$Upper\ limit = [3rd\ Quartile] + 1.5 \times [Interquartile\ range]$$

All results for the outlier analysis are included in the Supplementary Data Spreadsheet, Specimen Dat', listed as 'Within range' or 'Outlier'. Outlier specimens were not included in the phylogenetic cluster analysis, even if they are preserved in a suitable lithology.

Conodont elements were excluded from the dental tissue data set, since their homology is debated. Lack of an established consensus phylogeny for *Tullimonstrum*, *Mayomyzon*, and *Gilpichthys* in the tissue category 'vertebrate soft tissues' obliged us to omit these specimens from the phylogenetic cluster analysis. References for all consensus phylogenies are included in the main text.

H. Glossary

Diagenesis: The process of *postmortem* chemical and physical alteration of a carcass during initial sediment burial (early diagenesis) and deeper burial (late diagenesis).

Organic matter: The sum of all molecules based on organic carbon in a given fossil sample.

Nucleophiles: A molecule or ion that can act as an electron pair donor to form a new covalent bond with a reacting electron pair acceptor.

Functional groups: Small, distinct units in organic molecules that exhibit characteristic chemical properties.

Amide, amino acid, amino acid residue, peptide, protein: A peptide is an organic polymer consisting of condensed amino acids. The bond between two amino acids in a peptide is called a ‘peptide bond’ and can be chemically characterized as a *trans*-amide. All amino acids contain a terminal amine and carboxyl group, and only their characteristic amino acid residues differ. Resulting polymers (=peptides) contain an N-terminus (amine group) and a C-terminus (carboxyl group). Proteins differ from peptides in that they are larger, hierarchically arranged up to a quaternary superstructure, and post-translationally modified. Only unaltered peptides and proteins can be sequenced. During glycoxidation and lipoxidation, initial crosslinks almost exclusively involve the amino acid residues of lysine, arginine, cysteine, and histidine, as well as the N-terminal amine. Advanced crosslinking commonly also involves the peptide skeleton and consumes peptide bonds. Evidence for preserved *trans*-amides, and therefore unaltered peptide bonds, in fossils does not imply that attached amino acid residues are also unaltered. Crosslinking generates 3-dimensionally branching, highly stable polymers which, even if they preserve a few unaltered peptide bonds, are non-proteinaceous in nature (i.e., do not contain original secondary, tertiary, and quaternary structures).